(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 September 2004 (02.09.2004)

PCT

(10) International Publication Number
WO 2004/075029 A2

(51) International Patent Classification[7]: G06F

(21) International Application Number:
PCT/US2004/005172

(22) International Filing Date: 20 February 2004 (20.02.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/371,987 20 February 2003 (20.02.2003) US

(71) Applicant *(for all designated States except US)*: **MAILFRONTIER, INC.** [US/US]; 1801 Page Mill, Building F/G, Palo Alto, CA 94304 (US).

(72) Inventors: **WILSON, Brian, K.**; 425A Forest Avenue, Palo Alto, CA 94301 (US). **KOBLAS, David, A.**; 729 Anderson Drive, Los Altos, CA 94024 (US). **PENZIAS, Arno, A.**; 1960 Grant Avenue, San Francisco, CA 94133 (US).

(74) Agent: **VAN PELT, Lee**; Van Pelt & Yi LLP, Suite 200, 10050 N. Foothill Boulevard, Cupertino, CA 95014 (US).

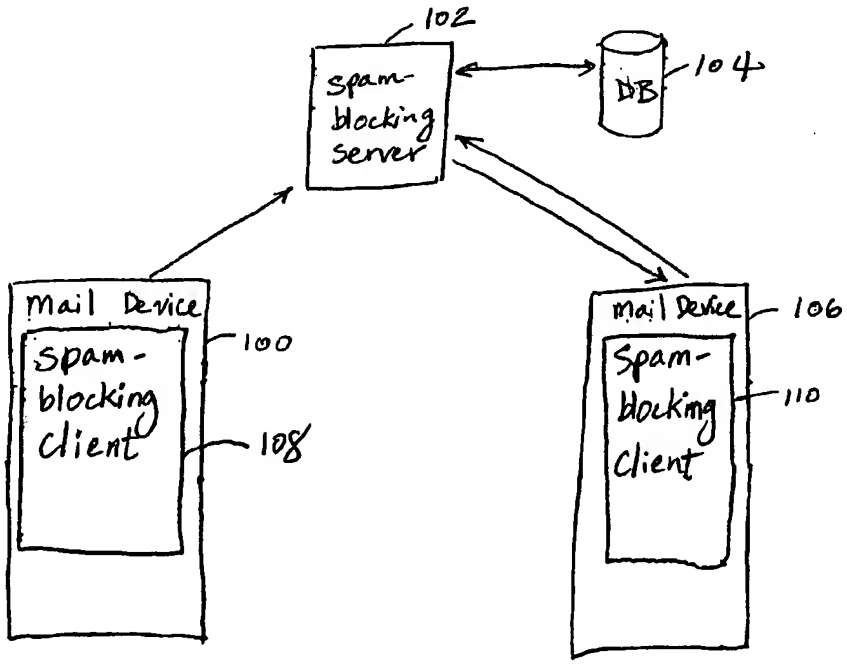
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: USING DISTINGUISHING PROPERTIES TO CLASSIFY MESSAGES



(57) Abstract: A system and method are disclosed for classifying a message. The method includes receiving the message, identifying in the message a distinguishing property; generating a signature using the distinguishing property; and comparing the signature to a database of signatures generated by previously classified messages.

WO 2004/075029 A2

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

USING DISTINGUISHING PROPERTIES TO CLASSIFY MESSAGES

CROSS REFERENCE TO RELATED APPLICATIONS

This application is related to co-pending U.S. Patent Application No. _____ (Attorney Docket No. MAILP002) entitled "MESSAGE IDENTIFICATION USING SUMMARY" filed concurrently herewith, which is incorporated herein by reference for all purposes.

FIELD OF THE INVENTION

The present invention relates generally to message classification. More specifically, a system and method for classifying messages that are junk email messages (spam) are disclosed.

BACKGROUND OF THE INVENTION

People have become increasingly dependent on email for their daily communication. Email is popular because it is fast, easy, and has little incremental cost. Unfortunately, these advantages of email are also exploited by marketers who regularly send out large amounts of unsolicited junk email (also referred to as "spam"). Spam messages are a nuisance for email users. They clog people's email box, waste system resources, often promote distasteful subjects, and sometimes sponsor outright scams.

There have been efforts to block spam using spam-blocking software in a collaborative environment where users contribute to a common spam knowledge base. For privacy and efficiency reasons, the spam-blocking software generally identifies spam messages by using a signature generated based on the content of the message. A relatively straightforward scheme to generate a signature is to first remove leading and trailing blank lines then compute a checksum on the remaining message body. However, spam senders (also referred to as "spammers") have been able to get around this scheme by embedding variations – often as random strings – in the messages so that the messages sent are not identical and generate different signatures.

Another spam-blocking mechanism is to remove words that are not found in the dictionary as well as leading and trailing blank lines, and then compute the checksum on the remaining message body. However, spammers have been able to circumvent this scheme by adding random dictionary words in the text. These superfluous words are sometimes added as white text on a white background, so that they are invisible to the readers but nevertheless confusing to the spam-blocking software.

The existing spam-blocking mechanisms have their limitations. Once the spammers learn how the signatures for the messages are generated, they can alter their message generation software to overcome the blocking mechanism. It would be desirable to have a way to identify messages that cannot be easily overcome even if the identification scheme is known. It would also be useful if any antidote to the identification scheme were expensive to implement or would incur significant runtime costs.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be readily understood by the following detailed description in conjunction with the accompanying drawings, wherein like reference numerals designate like structural elements, and in which:

Fig. 1 is a block diagram illustrating a spam message classification network according to one embodiment of the present invention.

Fig. 2 is a flowchart illustrating how to extract the distinguishing properties and use them to identify a message, according to one embodiment of the present invention.

Fig. 3 is a flowchart illustrating how a user classifies a message as spam according to one embodiment of the present invention.

Fig. 4 is a flowchart illustrating how the distinguishing properties are identified according to one embodiment of the present invention.

Fig. 5 is a flowchart illustrating the details of the email address identification step shown in Fig. 4.

DETAILED DESCRIPTION

It should be appreciated that the present invention can be implemented in numerous ways, including as a process, an apparatus, a system, or a computer readable medium such as a computer readable storage medium or a computer network wherein program instructions are sent over optical or electronic communication links. It should

be noted that the order of the steps of disclosed processes may be altered within the scope of the invention.

A detailed description of one or more preferred embodiments of the invention is provided below along with accompanying figures that illustrate by way of example the principles of the invention. While the invention is described in connection with such embodiments, it should be understood that the invention is not limited to any embodiment. On the contrary, the scope of the invention is limited only by the appended claims and the invention encompasses numerous alternatives, modifications and equivalents. For the purpose of example, numerous specific details are set forth in the following description in order to provide a thorough understanding of the present invention. The present invention may be practiced according to the claims without some or all of these specific details. For the purpose of clarity, technical material that is known in the technical fields related to the invention has not been described in detail so that the present invention is not unnecessarily obscured.

An improved system and method for classifying mail messages are disclosed. In one embodiment, the distinguishing properties in a mail message are located and used to produce one or more signatures. The signatures for junk messages are stored in a database and used to classify these messages. Preferably, the distinguishing properties include some type of contact information.

Fig. 1 is a block diagram illustrating a spam message classification network according to one embodiment of the present invention. The system allows users in the

network to collaborate and build up a knowledge base of known spam messages, and uses this knowledge to block spam messages. A spam message is first received by a mail device 100. The mail device may be a mail server, a personal computer running a mail client, or any other appropriate device used to receive mail messages. A user reads the message and determines whether it is spam.

If the message is determined to be spam, the spam-blocking client 108 on the mail device provides some indicia for identifying the message. In one embodiment, the indicia include one or more signatures (also referred to as thumbprints) based on a set of distinguishing properties extracted from the message. The signatures are sent to a spam-blocking server 102, which stores the signatures in a database 104. Different types of databases are used in various embodiments, including commercial database products such as Oracle databases, files, or any other appropriate storage that allow data to be stored and retrieved. In one embodiment, the database keeps track of the number of times a signature has been identified as spam by other users of the system. The database may be located on the spam-blocking server device, on a network accessible by server 102, or on a network accessible by the mail devices. In some embodiments, the database is cached on the mail devices and updated periodically.

When another mail device 106 receives the same spam message, before it is displayed to the user, spam-blocking client software 110 generates one or more signatures for the message, and sends the signatures along with any other query information to the spam-blocking server. The spam-blocking server looks up the

signatures in the database, and replies with information regarding the signatures. The information in the reply helps mail device 106 determine whether the message is spam.

Mail device 106 may be configured to use information from the spam-blocking server to determine whether the message is spam in different ways. For example, the number of times the message was classified by other users as spam may be used. If the number of times exceeds some preset threshold, the mail device processes the message as spam. The number and types of matching signatures and the effect of one or more matches may also be configured. For example, the message may be considered spam if some of the signatures in the signature set are found in the database, or the message may be determined to be spam only if all the signatures are found in the database.

Spammers generally have some motives for sending spam messages. Although spam messages come in all kinds of forms and contain different types of information, nearly all of them contain some distinguishing properties (also referred to as essential information) for helping the senders fulfill their goals. For example, in order for the spammer to ever make money from a recipient, there must be some way for the recipient to contact the spammer. Thus, some type of contact information is included in most spam, whether in the form of a phone number, an address, or a URL. Alternatively, certain types of instructions may be included. These distinguishing properties, such as contact information, instructions for performing certain tasks, stock ticker symbols, names of products or people, or any other information essential for the message, are extracted and used to identify messages. Since information that is not distinguishing is

discarded, it is harder for the spammers to alter their message generation scheme to evade detection.

It is advantageous that messages other than those sent by the spammer are not likely to include the same contact information or instructions. Therefore, if suitable distinguishing properties are identified, the risk of a false positive classification as spam can be diminished.

In some embodiments, spam-blocking server 102 acts as a gateway for messages. The server includes many of the same functions as the spam-blocking client. An incoming message is received by the server. The server uses the distinguishing properties in the messages to identify the messages, and then processes the messages accordingly.

Fig. 2 is a flowchart illustrating how to extract the distinguishing properties and use them to identify a message, according to one embodiment of the present invention. First, a message is received (200). The distinguishing properties in the message are identified (202), and one or more signatures are generated based on the distinguishing properties (204). The signatures are looked up in a database (206). If the signatures are not found in the database, then the system proceeds to process the message as a normal message, delivering the message or displaying it when appropriate (208). Otherwise, if matching signatures are found in the database, some appropriate action is taken accordingly (210). In an embodiment where the process takes place on a mail client, the action includes classifying the message as spam and moving it to an appropriate junk

folder. In an embodiment where the process takes place on a mail server, the action includes quarantining the message so it is recoverable by the administrator or the user.

Sometimes, a spam message is delivered to the user's inbox because an insufficient number of signature matches are found. This may happen the first time a spam message with a distinguishing property is sent, when the message is yet to be classified as spam by a sufficient number of users on the network, or when not enough variants of the message have been identified. The user who received the message can then make a contribution to the database by indicating that the message is spam. In one embodiment, the mail client software includes a "junk" button in its user interface. The user can click on this button to indicate that a message is junk. Without further action from the user, the software automatically extracts information from the message, submits the information to the server, and deletes the message from the user's inbox. In some embodiments, the mail client software also updates the user's configurations accordingly. For instance, the software may add the sender's address to a blacklist. The blacklist is a list of addresses used for blocking messages. Once an address is included in the blacklist, future messages from that address are automatically blocked.

Fig. 3 is a flowchart illustrating how a user classifies a message as spam according to one embodiment of the present invention. A spam message is received by the user (300). The user selects the message (302), and indicates that the message is junk by clicking on an appropriate button or some other appropriate means (304). The software identifies the distinguishing properties in the message (306), and generates a set of signatures based on the distinguishing properties (308). The signatures are then

submitted to the database (310). Thus, matching signatures can be found in the database for messages that have similar distinguishing properties. In some embodiments, the mail client software then updates the user's configurations based on the classification (312). In some embodiments, the sender's address is added to a blacklist. The message is then deleted from the user's inbox (314).

Fig. 4 is a flowchart illustrating how the distinguishing properties are identified according to one embodiment of the present invention. Since most spammers would like to be contacted somehow, the messages often include some sort of contact information, such as universal resource locators (URL's), email addresses, Internet protocol (IP) addresses, telephone numbers, as well as physical mailing addresses. In this embodiment, the distinguishing properties of the message include contact information.

The message is preprocessed to remove some of the non-essential information (400), such as spaces, carriage returns, tabs, blank lines, punctuations, and certain HTML tags (color, font, etc.).

Distinguishing properties are then identified and extracted from the message. Since spammers often randomly change the variable portions of URL's and email addresses to evade detection, the part that is harder to change – the domain name – is included in the distinguishing properties while the variable portions are ignored. The domain name is harder to change because a fee must be paid to obtain a valid domain name, making it less likely that any spammer would register for a large number of domain names just to evade detection. The software scans the preprocessed message to

identify URL's in the text, and extracts the domain names from the URL's (402). It also processes the message to identify email addresses in the text and extracts the domain names embedded in the email addresses (404).

Telephone numbers are also identified (406). After preprocessing, phone numbers often appear as ten or eleven digits of numbers, with optional parentheses around the first three digits, and optional dashes and spaces between the numbers. The numbers are identified and added to the distinguishing properties. Physical addresses are also identified using heuristics well known to those skilled in the art (408). Some junk messages may contain other distinguishing properties such as date and location of events, stock ticker symbols, etc. In this embodiment, these other distinguishing properties are also identified (410). It should be noted that the processing steps are performed in different order in other embodiments. In some embodiments, a subset of the processing steps is performed.

Fig. 5 is a flowchart illustrating the details of the email address identification step shown in Fig. 4. First, the message is scanned to find candidate sections that include top-level domain names (500). The top-level domain refers to the last section of an address, such as .com, .net, .uk, etc. An email address includes multiple fields separated by periods. The top-level domain determines which fields form the actual domain name, according to well-known standards. For example, the address user1@server1.mailfrontier.com has a domain name that includes two fields (mailfrontier.com), while as user2@server1.mailfrontier.co.uk has a domain name that includes three fields (mailfrontier.co.uk). Thus, the top-level domain in a candidate

section is identified (502), and the domain name is determined based on the top-level domain (504).

The presence of any required characters (such as @) is checked to determine whether the address is a valid email addresses (506). If the address does not include the require characters, it is invalid and its domain name should be excluded from the distinguishing properties (514). If the required characters are included in the address, any forbidden characters (such as commas and spaces) in the address are also checked (508). If the address includes such forbidden characters, it is invalid and its domain name may be excluded from the distinguishing properties (514).

Sometimes, spammers embed decoy addresses – fake addresses that have well-known domain names – in the messages, attempting to confuse the spam-blocking software. In some embodiments, the decoy addresses are not included in the distinguishing properties. To exclude decoy addresses, an address is checked against a white list of well-known domains (510), and is excluded from the distinguishing properties if a match is found (514). If the address is not found in the white list, it belongs to the distinguishing properties (512).

In some embodiments, a similar process is used to identify URL's. The domain names of the URL's are extracted and included in the distinguishing properties, and decoy URL's are discarded. Sometimes, spammers use numerical IP addresses to hide their domain names. By searching through the message for any URL that has the form http://x.x.x.x where the x's are integers between 0-255, these numerical IP addresses are

identified and included in the distinguishing properties. More crafty spammers sometimes use obscure forms of URL's to evade detection. For example, binary numbers or a single 32 bit number can be used instead of the standard dotted notation. Using methods well-known to those skilled in the art, URL's in obscure forms can be identified and included in the distinguishing properties. In some embodiments, physical addresses, events, and stock quotes are also identified.

Once the distinguishing properties have been identified, the system generates one or more signatures based on the distinguishing properties and sends the signatures to the database. The signatures can be generated using a variety of methods, including compression, expansion, checksum, or any other appropriate method. In some embodiments, the data in the distinguishing properties is used directly as signatures without using any transformation. In some embodiments, a hash function is used to produce the signatures. Various hash functions are used in different embodiments, including MD5 and SHA. In some embodiments, the hash function is separately applied to every property in the set of distinguishing properties to produce a plurality of signatures. In one embodiment, any of the distinguishing properties must meet certain minimum byte requirement for it to generate a corresponding signature. Any property that has fewer than a predefined number of bytes is discarded to lower the probability of signature collisions.

The generated signatures are transferred and stored in the database. In one embodiment, the signatures are formatted and transferred using extensible markup language (XML). In some embodiments, the signatures are correlated and the

relationships among them are also recorded in the database. For example, if signatures from different messages share a certain signature combination, other messages that include the same signature combination may be classified as spam automatically. In some embodiments, the number of times each signature has been sent to the database is updated.

Using signatures to identify a message gives the system greater flexibility and allows it to be more expandable. For example, the mail client software may only identify one type of distinguishing property in its first version. In later versions, new types of distinguishing properties are added. The system can be upgraded without requiring changes in the spam-blocking server and the database.

An improved system and method for classifying a message have been disclosed. The system identifies the distinguishing properties in an email message and generates one or more signatures based on the distinguishing properties. The signatures are stored in a database and used by spam-blocking software to effectively block spam messages.

Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. It should be noted that there are many alternative ways of implementing both the process and apparatus of the present invention. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

WHAT IS CLAIMED IS:

CLAIMS

1. A method for classifying a message comprising:

receiving the message;

identifying in the message a distinguishing property;

generating a signature using the distinguishing property; and

comparing the signature to a database of signatures generated by previously classified messages.

2. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes contact information.

3. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes contact information; and the contact information includes an email address.

4. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes contact information; and the contact information includes a telephone number.

5. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes contact information; and the contact information includes a universal resource locator (URL).

6. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes contact information; and the contact information includes an Internet Protocol (IP) address.

7. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes contact information; and the contact information includes a domain name.

8. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes a name.

9. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes a stock ticker symbol.

10. A method for classifying a message as recited in Claim 1 wherein the distinguishing property includes instructions for performing a task.

11. A method for classifying a message as recited in Claim 1 further including determining whether the signature exists in a database of previously stored signatures.

12. A method for classifying a message as recited in Claim 1 further including determining whether the signature exists in a database of previously stored signatures; and in the event that the signature does not exist in the database, adding the signature to the database.

13. A method for classifying a message as recited in Claim 1 further including adding the signature to a database.

14. A method for classifying a message as recited in Claim 1 further including adding the signature to a database; wherein the database tracks the number of times the message has been identified as junk message.

15. A method for classifying a message as recited in Claim 1 further including adding the signature to a database; wherein the database tracks the number of times the message has been identified as junk message and the database is located on a mail device.

16. A method for classifying a message as recited in Claim 1 further including adding the signature to a database; wherein the database is located on a server.

17. A method for classifying a message as recited in Claim 1 further including adding the signature to a database; wherein the database is located on a network accessible by a spam-blocking server.

18. A method for classifying a message as recited in Claim 1 wherein the signature is transferred to a database using an XML based protocol.

19. A method for classifying a message as recited in Claim 1 wherein generating the signature includes performing a hash function on the distinguishing property.

20. A method for classifying a message as recited in Claim 1 wherein generating the signature includes performing a hash function on the distinguishing property; and the hash function is a SHA function.

21. A method for classifying a message as recited in Claim 1 wherein generating the signature includes performing a hash function on the distinguishing property; and the hash function is an MD5 function.

22. A method for classifying a message as recited in Claim 1 wherein generating the signature includes performing a hash function on the distinguishing property and the hash function is a checksum function.

23. A method for classifying a message as recited in Claim 1 wherein the signature is generate for distinguishing properties that meet a minimum byte requirement.

24. A method for classifying a message as recited in Claim 1 wherein identifying the distinguishing property comprises preprocessing the message to remove non-essential information.

25. A method for classifying a message as recited in Claim 1 wherein identifying the distinguishing property comprises processing the message to exclude decoy information.

26. A computer program product for classifying a message, the computer program product being embodied in a computer readable medium and comprising computer instructions for:

receiving the message;

identifying in the message an distinguishing property;

generating a signature using the distinguishing property; and

comparing the signature to a database of signatures generated by previously classified messages.

27. A system for classifying a message comprising:

an interface configured to receive a message; and

a processor configured to:

identify in the message a distinguishing property;

generate a signature using the distinguishing property; and

compare the signature to a database of signatures generated by previously classified messages.

28. A method for classifying a message comprising:

receiving the message;

receiving a classification of the message;

identifying in the message a distinguishing property;

generating a signature using the distinguishing property;

submitting the signature to a database.

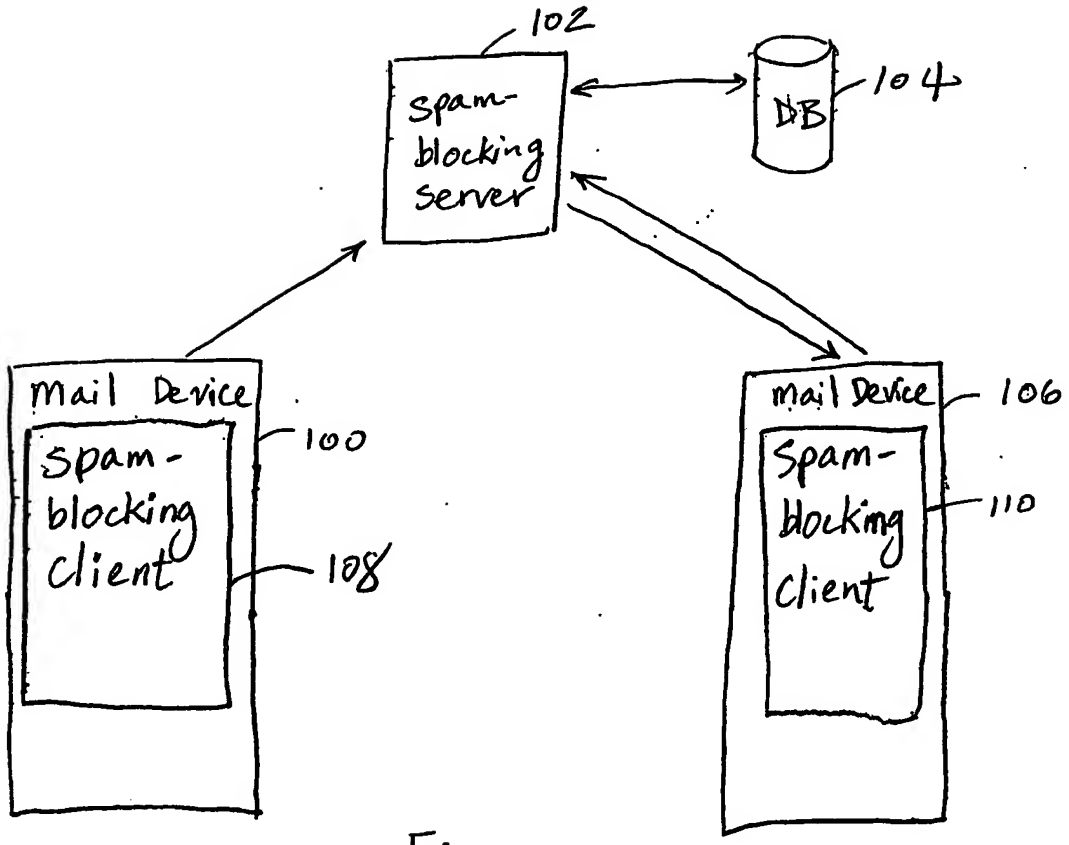

102
spam-blocking server
DB
104
mail Device
spam-blocking client
100
108
mail Device
spam-blocking client
106
110

Fig. 1

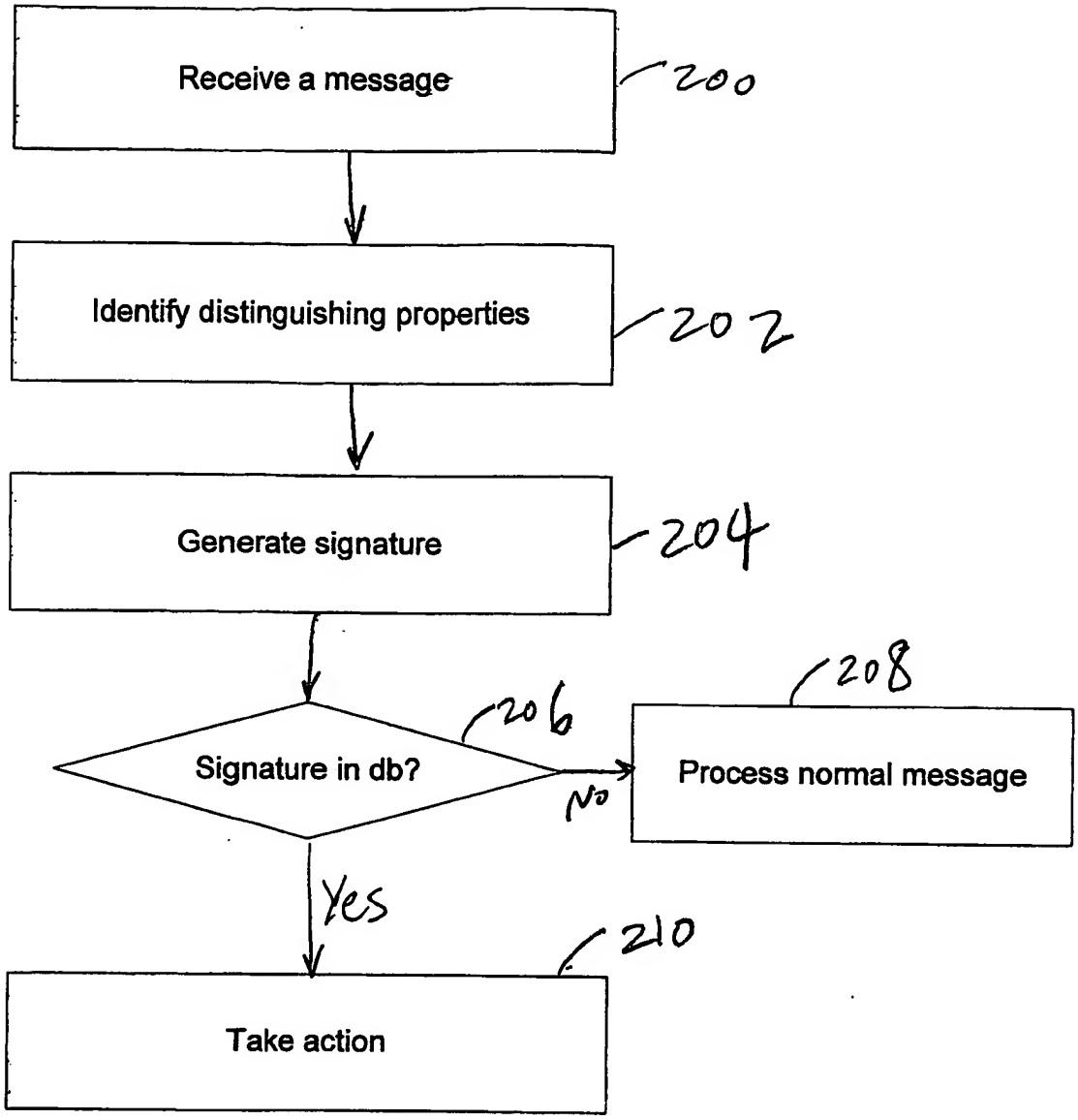

Receive a message
200
Identify distinguishing properties
202
Generate signature
204
Signature in db?
206
No
Process normal message
208
Yes
Take action
210

Fig. 2

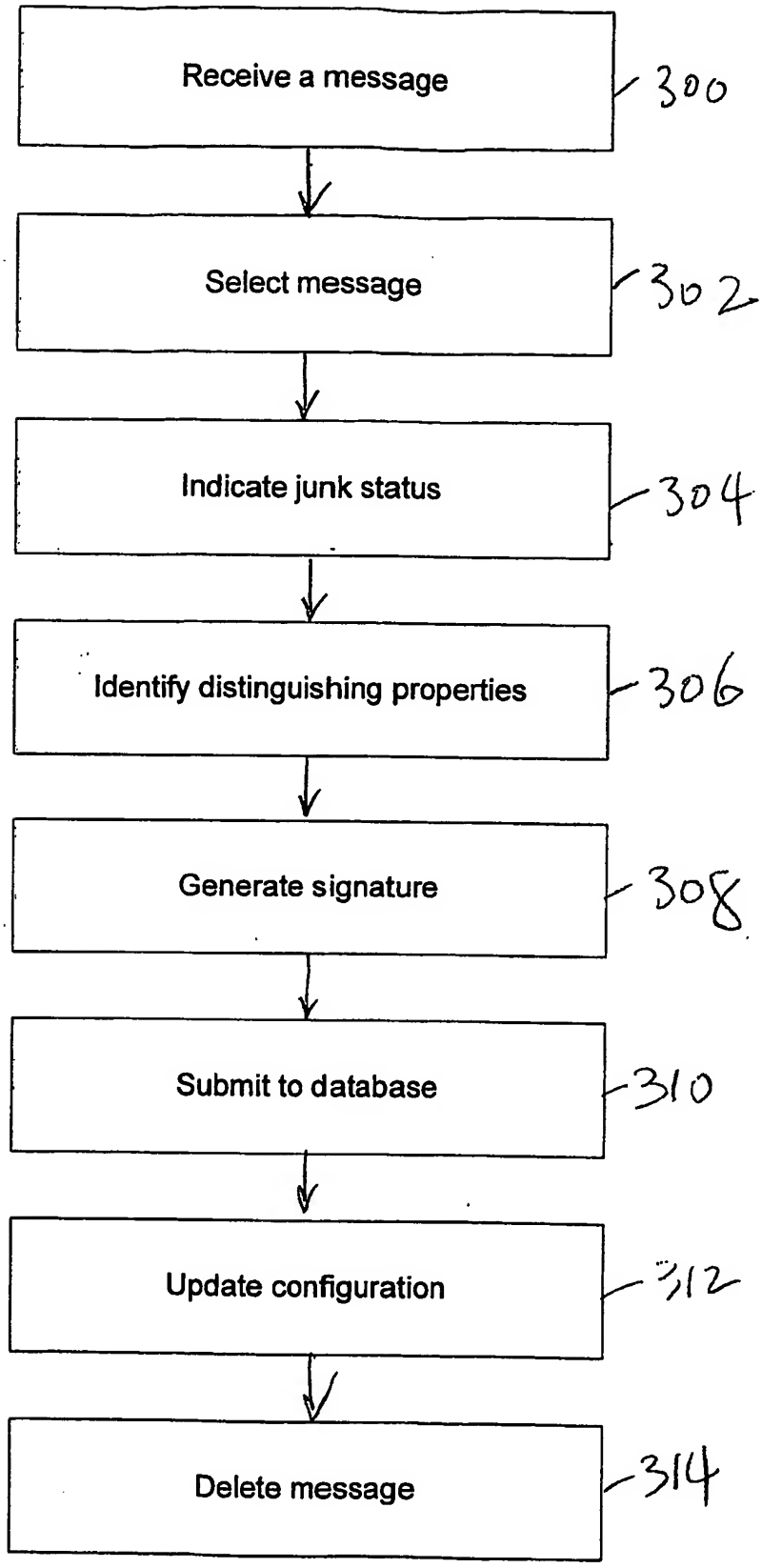

Receive a message
300
Select message
302
Indicate junk status
304
Identify distinguishing properties
306
Generate signature
308
Submit to database
310
Update configuration
312
Delete message
314

Fig. 3

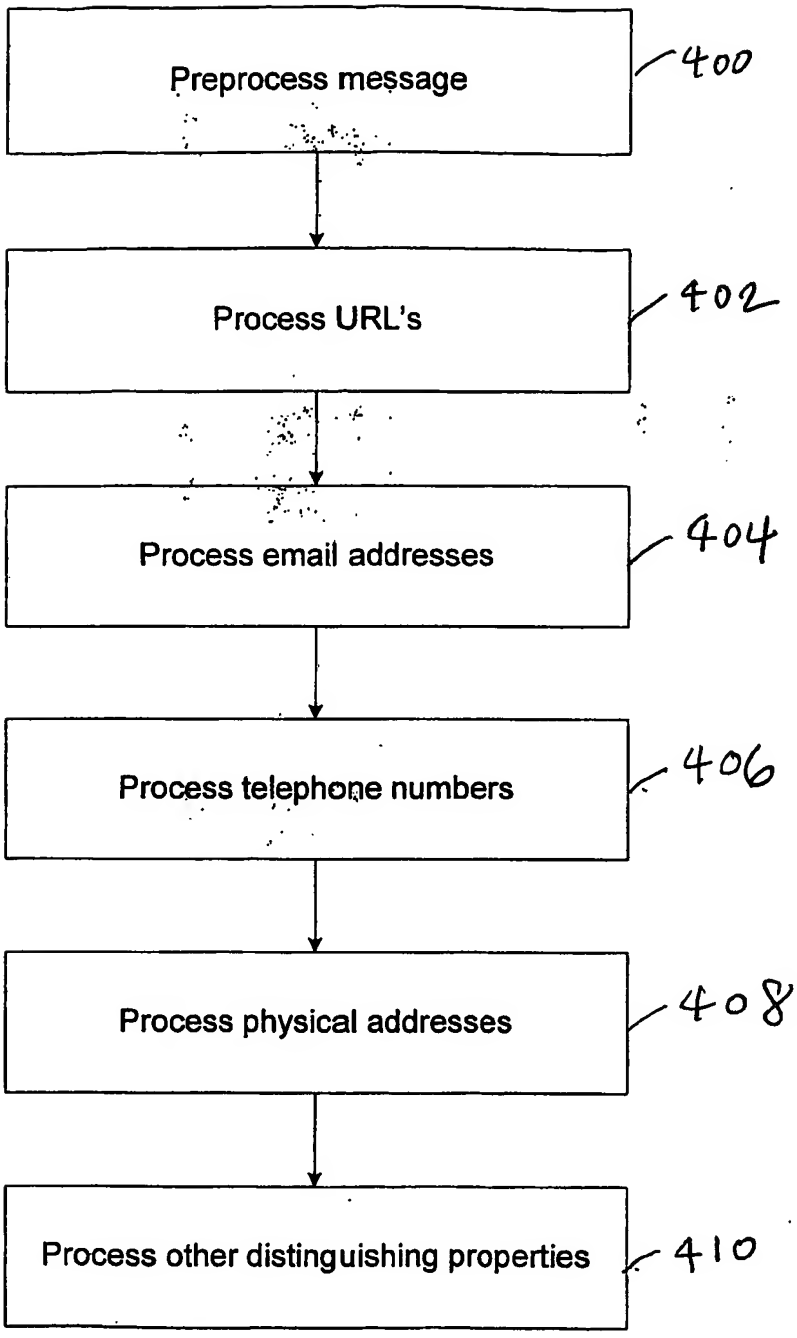

Preprocess message
400
Process URL's
402
Process email addresses
404
Process telephone numbers
406
Process physical addresses
408
Process other distinguishing properties
410

Fig. 4

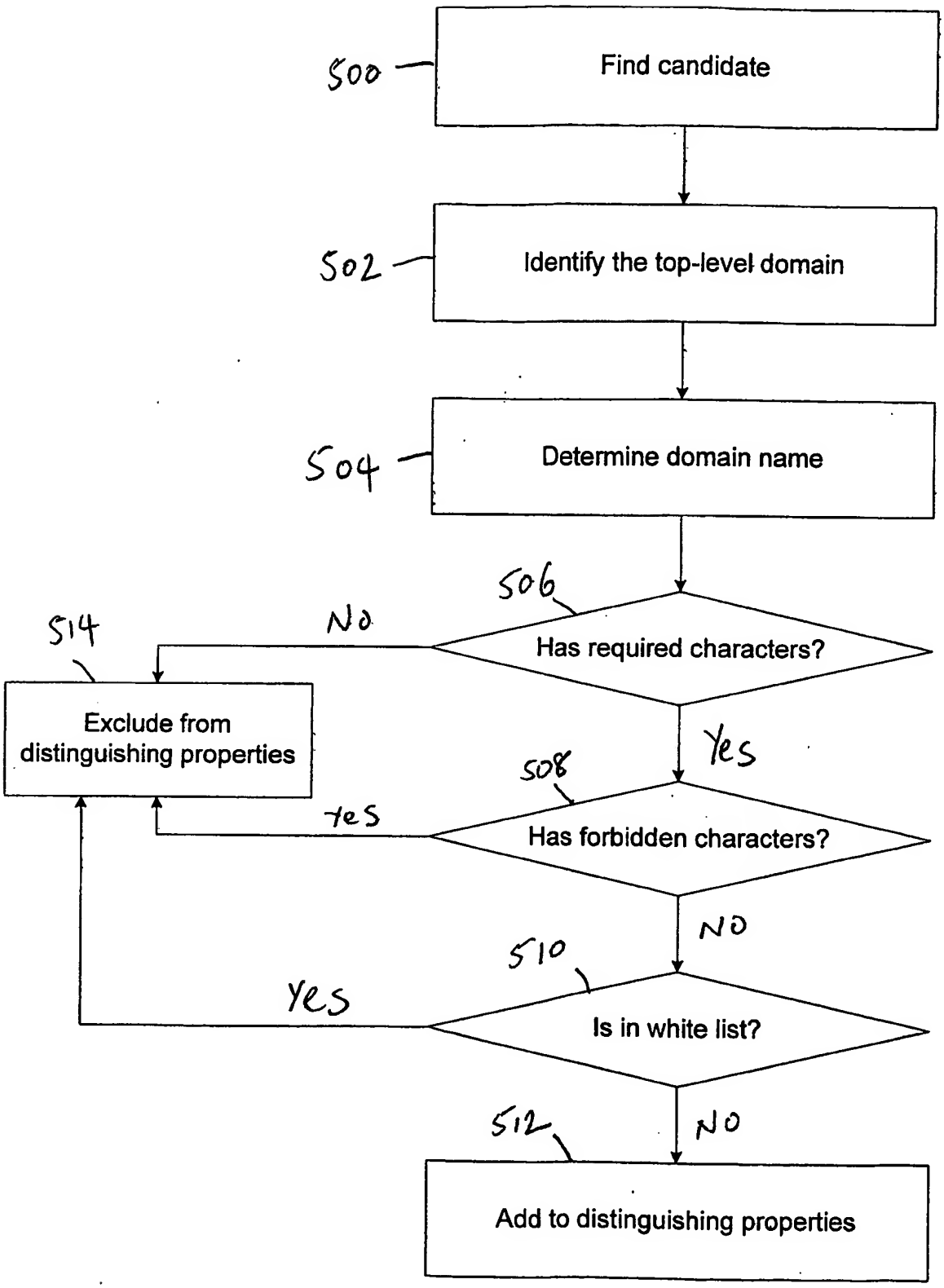

500
Find candidate
502
Identify the top-level domain
504
Determine domain name
506
Has required characters?
No
514
Exclude from distinguishing properties
Yes
508
Has forbidden characters?
Yes
NO
510
Is in white list?
Yes
NO
512
Add to distinguishing properties

Fig. 5